DOCUMENT RESUME

ED 186 494 TM 800 349

AUTHOR Vermillion, James E.
TITLE Regression Artifacts in Nonequivalent Control Group Designs: An Empirical Investigation of Bias in ANCOVA and Matching Designs.
PUB DATE Apr 80
NOTE 24p.: Paper presented at the Annual Meeting of the American Educational Research Association (64th, Boston, MA, April 7-11, 1980).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Analysis of Covariance; Control Groups; Error of Measurement; Experimental Groups; Foreign Countries; *Matched Groups; Preschool Education; *Quasiexperimental Design; Regression (Statistics); Research Design; *Sampling; *Statistical Bias; Statistical Significance
IDENTIFIERS Mexico (Mexico City)

ABSTRACT

The presence of artifactual bias in analysis of covariance (ANCOVA) and in matching nonequivalent control group (NECG) designs was empirically investigated. The data set was obtained from a study of the effects of a television program on children from three day care centers in Mexico in which the subjects had been randomly selected within centers. However, differences in the scores of the children between centers were sufficient to provide a more able and a less able group for subsamples in an artificial quasiexperimental study of the same variables. With these subsamples, the same treatment effects were estimated using each NECG design. Unbiased and NECG estimators were then compared for each design, using the one-tailed binomial test to seek the presence of a systematic artifactual bias. A significant bias underrating the group initially lower on the pretest measures was found in both designs. The use of modern multivariate analysis designs was recommended as a better alternative, but experimental studies of these designs were also recommended. The advantage of randomization and true experimental design was pointed out. (Author/CTM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED186494

U S DEPARTMENT OF HEALTH
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

REGRESSION ARTIFACTS IN NONEQUIVALENT CONTROL GROUP DESIGNS: AN EMPIRICAL INVESTIGATION OF BIAS IN 'ANCOVA' AND 'MATCHING' DESIGNS

James E. Vermillion

Boston University School of Nursing

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
J. Vermillion

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM800349

A paper presented to the 1980 Annual Meeting of the American Educational Research Association, Boston, Ma.

Attempts to adjust for initial differences between experimental and control groups have often led evaluators to choose nonequivalent control group (NECG) analysis of covariance (ANCOVA) or 'matching' designs. In particular such methodology has sometimes been employed in the evaluation of compensatory education programs. Claims have been made that inherent in such designs is an artifactual bias which systematically underrates whichever group is initially lower.

Almost all traditional regression models make assumptions regarding the distribution of error, in particular independence of error among subjects. These assumptions serve as a foundation for the tests of statistical significance of treatment methods. Use of intact groups frequently results in violation of this assumption. In most models the error among subjects within the intact group contains systematic variance related to variables unspecified in the model, but common to all members of the intact but not to members of the control group. Although recently techniques using Bayesian statistical methods which do not assume this independence of error among subjects (e.g., Joreskog, 1970) have become popular, most classical models in the literature make the assumption of non-systematic error among subjects within each group. This assumption necessitates strict random assignment to groups in experiments so as not to vitiate the statistical tests of significance. In the social sciences, random group assignment is often almost impossible given the cost, ethical, and political considerations involved. Hence intact groups are often used in evaluation and research designs. Control groups are generally in these instances drawn from the general population, and by the use of 'matching' or by the use of ANCOVA a comparison technique is available. Problems with 'matching' and ANCOVA have been associated with this methodology since at least 1953 by Cochran in his article "Matching in Analytical Studies", and 1960 by Lord in his article "Large Scale Covariance Analysis when the Control Variable is Fallible". The basic difficulty theorized was that control in this manner allows a regression of scores to the general population mean in the control group, but for the intact group those systematic forces which caused the group to be different from the general population in the first place continue to act to keep the scores in this group from regressing as strongly toward the general population mean. Indeed, Campbell has suggested that often the regression in the intact group is stronger in the opposite direction from, rather than toward, the population mean underrating the initially lower group (Campbell & Erlebacher, 1970).

The Ohio/Westinghouse study of the effects of the 1968 summer Head Start program employed a 'matching' design and used analysis of covariance to guard against bias due to initial group differences. This study concluded that the Head Start program was totally ineffective. Re-examination of the data with different designs has raised questions regarding the appropriateness of the conclusions drawn in the Ohio/Westinghouse study (e.g., Magidson, 1977). It seems that a bias as described by Campbell and Stanley (Campbell & Stanley, 1963) and others might have influenced the investigators in their conclusions in this evaluation.

Campbell (e.g., Campbell & Boruch, 1975) has described this bias as directly proportional to the initial differences between the groups and in a direction which underrates the group which is initially lower on premeasures. Cronbach and others (Cronbach et al., 1977) have

indicated that the magnitude and direction of the bias vary as a function of the correlation between the actual covariate(s) and a function which discriminates between the groups. The question addressed by the present study was:

Do ANCOVA and 'matching' NECG quasiexperimental designs produce an artifactual bias which underrates whichever group is initially lower on premeasures?

In particular, two specific research questions were addressed to answer the above question:

(1) Does the ANCOVA NECG design create an artifactual bias which underrates whichever group is initially lower?

(2) Does the 'matching' NECG design create an artifactual bias which underrates whichever group is initially lower?

DATA SOURCE

The data set used in this study was that collected in a study of the effect of the television program Plaza Sesamo, a Mexican adaptation of the American-produced educational program Sesame Street.* The data provide test results of 221 children from three day care centers around Mexico City on nine different tests at three points in time. Subjects within centers were randomly assigned to experimental and control groups. Two of the centers were sufficiently different on pretest scores to provide a 'more able' and a 'less able' center as a basis for NECG designs in subsampling. The study producing the data set was very well controlled, with experimental group children watching Plaza Sesamo and control children watching cartoons over the six months of the study. Consistently significant effects in favor of the Plaza Sesamo treatment were found.

The nine tests for which scores were available were General Knowledge, Numbers, Relations, Embedded Figures, Parts of the Whole, Ability to Classify, Letters and Words, Classification Skills, and Oral Comprehension. Descriptions of these nine tests were given by the original investigators (Diaz-Guerro & Holtzman, 1973, p.4-6). Psychometric properties of the tests were also reported by the original investigators. Test-retest stability was reported as satisfactory (.41 to .59) for the General Knowledge, Numbers, Embedded Figures, and Oral Comprehension tests and as low (.08 to .25) for the remaining tests (Diaz-Guerrero and Holtzman, 1976, p. 147). These psychometric properties were not directly relevant to this investigation since emphasis here was not on generalizability to constructs measured by these instruments but rather on estimates of effects produced in ANCOVA and 'matching' NECG designs employing subsets of the data.

---

* The author wishes to acknowledge and thank R. Diaz-Guerrero, President of the Instituto Nacional de Ciencias del Comportamiento y de la Actitud Publica, A.C., Mexico City, and Jay Magidson of Abt Associates, Inc., Cambridge, Massachusetts, who have made available the data for this investigation.

The center with pretest scores between those of the other two centers was excluded from this analysis since if any artifactual bias was expected to appear, it was most likely to appear when the experimental and control groups were most different on premeasures. Also, since lowest pretest-posttest gains were found among three-year-olds, this group of children was also excluded from these analyses. Finally, only children with complete data on all nine tests at all three time points were retained for the analysis.

RESEARCH DESIGN

The specific research questions addressed in this study were:

(1) Does the ANCOVA NECG design create an artifactual bias which underrates whichever group is initially lower?

(2) Does the 'matching' NECG design create an artifactual bias which underrates whichever group is initially lower?

Separate analyses were performed to answer each question. Table 1 shows the direction of bias predicted for both the ANCOVA and 'matching' designs in the case of the 'more able' and of the 'less able' experimental group.

TABLE 1

PREDICTION OF BIAS DIRECTION IN NECG QUASIEXPERIMENTAL DESIGNS

Technique	Experimental Group	Control Group	Predicted Bias
ANCOVA	'less able'	'more able'	underrate treatment
ANCOVA	'more able'	'less able'	overrate treatment
'matching'	'less able'	'more able'	underrate treatment
'matching'	'more able'	'less able'	overrate treatment

ANCOVA NECG Designs

To determine whether ANCOVA NECG designs produce artifactual bias, unbiased estimators of treatment effect were obtained for each of the nine tests using (1) posttest adjusted for pretest by ANCOVA, and then (2) midtest adjusted for pretest by ANCOVA. Treatment effect was defined as the difference between the adjusted means of the experimental and the control groups. By subsampling, a quasiexperiment was simulated with the experimental group selected from the 'less able' center and with the control group drawn from the 'more able' center. Estimators of the treatment effect were obtained in this simulated 'less able' experimental group NECG design for both posttest-pretest adjustment and midtest-pretest adjustment. These NECG design estimators were compared with the unbiased ANCOVA estimators for the posttest-pretest and midtest-pretest conditions under a binomial test of null hypothesis #1:

p(unbiased effect estimator > NECG design effect estimator) =
p(unbiased effect estimator < NECG design effect estimator) = 0.5

against the one-tailed alternate hypothesis:

p(unbiased effect estimator > NECG design effect estimator) >
p(unbiased effect estimator < NECG design effect estimator)

(with an alpha level of .05). The one-tailed alternative seemed reasonable since the literature suggested that any bias would tend to underrate whichever group was initially lower, in this case the experimental group, leading to a deflation of treatment effect.

Next, simulations of 'more able' experimental group NECG ANCOVA designs were produced by subsampling of the experimental group from the 'more able' center and of the control group from the 'less able' center. Again, estimators of treatment effect were obtained for both posttest-pretest and midtest-pretest adjustments. These 'more able' experimental group NECG design estimators were then compared with the unbiased ANCOVA estimators for the posttest-pretest and midtest-pretest adjustments, under a binomial test of null hypothesis #2:

p(unbiased effect estimator > NECG design effect estimator) =
p(unbiased effect estimator < NECG design effect estimator) = 0.5

against the one-tailed alternate hypothesis:

p(unbiased effect estimator > NECG design effect estimator) <
p(unbiased effect estimator < NECG design effect estimator)

(with an alpha level of .05). Here, since the experimental group was 'more able' and the control group 'less able' in the NECG design, the bias in that design was expected to underrate the control group, artifactually inflating the treatment effect.

'Matching' NECG Designs

To investigate whether 'matching' designs produced artifactual bias, again unbiased estimators of treatment effect were obtained. Subjects from the 'less able' center's experimental group were matched with subjects from the 'more able' center's control group on age, sex, and pretest scores. Estimators of treatment effect in this 'less able' experimental group NECG design were obtained and compared with the unbiased effect estimators for both the posttest and the midtest under a binomial test of null hypothesis #3:

p(unbiased effect estimator > NECG design effect estimator) =
p(unbiased effect estimator < NECG design effect estimator) = 0.5

This was tested against the one-tailed alternate hypothesis:

p(unbiased effect estimator > NECG design effect estimator) >
p(unbiased effect estimator < NECG design effect estimator)

(with an alpha level of .05). Again with the 'less able' experimental group design it was expected that the treatment effect would be

underrated.

Finally, a 'more able' experimental group NECG design was simulated by selecting subjects from the 'more able' center's experimental group and their matches (again on age, sex, and pretest scores) from the 'less able' center's control group. Estimators of treatment effect in this NECG design were obtained and compared with the unbiased estimators of effect under the binomial test of null hypothesis #4:

$$p(\text{unbiased effect estimator} > \text{NECG design effect estimator}) = p(\text{unbiased effect estimator} < \text{NECG design effect estimator}) = 0.5$$

against the one-tailed alternate hypothesis:

$$p(\text{unbiased effect estimator} > \text{NECG design effect estimator}) < p(\text{unbiased effect estimator} < \text{NECG design effect estimator})$$

(with an alpha level of .05). Since the experimental group was 'more able' the bias was expected to underrate the control group and inflate the estimator of treatment effect.

It was assumed that directions of differences between experimental and control group means and adjusted means were independent among the nine tests within each design and among designs. Because this experiment employed only one empirical data set to test mathematical expectations, the uniqueness of this data set allows only limited generalizability. The focus of this investigation was to demonstrate the contrast in conclusions which would have been reached had the methodology employed in this particular study been any of these NECG quasiexperimental designs in place of the true experimental design. The findings of this investigation provide evidence in this one instance of distortion resulting from artifactual biases related to use of particular quasiexperimental research methodologies employed.

FINDINGS

To answer the research question "Does the ANCOVA NECG design produce an artifactual bias underrating whichever group is initially lower?", first an ANCOVA was performed using the data from both the experimental and control groups in the 'more able' and 'less able' centers. Estimators of effect were determined first with the posttest adjusted for the pretest. The full data set estimators obtained here were the basis for later comparisons with estimators from NECG designs drawn from subsets of the data. Table 2 gives the adjusted means of the posttest, adjusted for pretest, the estimate of treatment effect (experimental adjusted mean minus the control adjusted mean), and the level of significance of difference between experimental and control group adjusted means (when $p < .05$).

TABLE 2

MEANS FOR POSTTEST ADJUSTED BY COVARIANCE FOR PRETEST IN THE RANDOMIZED GROUPS EXPERIMENTAL DESIGN

	Experimental (n = 39)	Control (n = 38)	Effect	P
General Knowledge	29.71	24.58	5.13	<.01
Numbers	34.36	27.35	7.01	<.01
Relations	5.74	5.11	0.63	
Embedded Figures	21.34	19.58	1.76	<.01
Parts of the Whole	6.20	5.40	0.80	
Ability to Classify	8.67	6.01	2.66	<.01
Letters and Words	16.92	12.50	4.42	<.01
Classification Skills	13.73	11.40	2.33	<.01
Oral Comprehension	25.53	22.45	3.08	<.01

Next, unbiased estimators of effect for the midtest, adjusted for pretest, were determined, as shown in Table 3.

TABLE 3

MEANS FOR MIDTEST ADJUSTED BY COVARIANCE FOR PRETEST IN THE RANDOMIZED GROUPS EXPERIMENTAL DESIGN

	Experimental (n = 39)	Control (n = 38)	Effect	P
General Knowledge	23.29	22.86	0.43	
Numbers	27.24	24.54	2.70	<.05
Relations	5.17	4.63	0.54	
Embedded Figures	18.50	17.39	1.11	
Parts of the Whole	5.04	4.25	0.79	
Ability to Classify	6.66	5.77	0.89	
Letters and Words	14.07	12.16	1.91	
Classification Skills	10.04	9.09	0.95	
Oral Comprehension	20.92	20.03	0.89	

To test whether artifactual bias was created in the 'less able' experimental group NECG ANCOVA design, such an NECG design was simulated by selecting the experimental group from the 'less able' center and the control group from the 'more able' center. Estimators of treatment effect were obtained in this NECG design for the posttest adjusted by ANCOVA for the pretest, as shown in Table 4.

TABLE 4

MEANS FOR POSTTEST ADJUSTED BY COVARIANCE FOR PRETEST
IN THE 'LESS ABLE' EXPERIMENTAL GROUP NECG ANCOVA DESIGN

	Experimental (n = 16)	Control (n = 22)	Effect	P
General Knowledge	29.18	25.50	3.68	
Numbers	33.44	27.95	5.49	<.05
Relations	5.06	5.77	-0.71	
Embedded Figures	21.08	19.58	1.50	
Parts of the Whole	5.96	5.61	0.35	
Ability to Classify	8.51	6.45	2.06	
Letters and Words	16.90	12.25	4.65	<.01
Classification Skills	12.98	11.57	1.41	
Oral Comprehension	24.70	23.40	1.30	

Estimators of the treatment effect on the midtest, adjusted for the pretest, were next determined for the 'less able' experimental group ANCOVA NECG design, as shown in Table 5.

TABLE 5

MEANS FOR MIDTEST ADJUSTED BY COVARIANCE FOR PRETEST
IN THE 'LESS ABLE' EXPERIMENTAL GROUP NECG ANCOVA DESIGN

	Experimental (n = 16)	Control (n = 22)	Effect	P
General Knowledge	21.90	24.61	-2.71	<.01
Numbers	26.41	24.93	1.48	
Relations	4.53	4.70	-0.17	
Embedded Figures	17.99	17.59	0.40	
Parts of the Whole	4.22	4.20	0.02	
Ability to Classify	5.95	6.40	-0.45	
Letters and Words	13.38	13.78	-0.40	
Classification Skills	8.49	9.56	-1.07	
Oral Comprehension	20.53	22.07	-1.54	

To test whether an artifactual bias, significantly underrating the experimental effect in the 'less able' experimental group NECG ANCOVA design was present, a binomial test of null hypothesis #1 was performed (see Table 6).

TABLE 6

COMPARISON OF 'LESS ABLE' EXPERIMENTAL GROUP NECG ANCOVA DESIGN ESTIMATORS OF EFFECTS WITH UNBIASED EFFECTS ESTIMATORS

	Unbiased Estimator	NECG Design Estimator	Unbiased Estimator > NECG Design Estimator
POSTTEST-PRETEST			
General Knowledge	5.13	3.68	yes
Numbers	7.01	5.49	yes
Relations	0.63	-0.71	yes
Embedded Figures	1.76	1.50	yes
Parts of the Whole	0.80	0.35	yes
Ability to Classify	2.66	2.06	yes
Letters and Words	4.42	4.65	no
Classification Skills	2.33	1.41	yes
Oral Comprehension	3.08	1.30	yes
MIDTEST-PRETEST			
General Knowledge	0.43	-2.71	yes
Numbers	2.70	1.48	yes
Relations	0.54	-0.17	yes
Embedded Figures	1.11	0.40	yes
Parts of the Whole	0.79	0.02	yes
Ability to Classify	0.89	-0.45	yes
Letters and Words	1.91	-0.40	yes
Classification Skills	0.95	-1.07	yes
Oral Comprehension	0.89	-1.54	yes

The binomial test yielded a $p < .001$ leading to a rejection of the null hypothesis and an acceptance of the alternate hypothesis that in 'less able' experimental group NECG ANCOVA designs a bias underrating the treatment effect appeared.

To test whether artifactual bias was created in the 'more able' experimental group NECG ANCOVA design, such a design was simulated by selecting the experimental group from the 'more able' center and the control group from the 'less able' center. Again estimators of treatment effect were obtained in this NECG design for when the posttest was adjusted by ANCOVA for the pretest, as shown in Table 7.

TABLE 7

MEANS FOR POSTTEST ADJUSTED BY COVARIANCE FOR PRETEST
IN THE 'MORE ABLE' EXPERIMENTAL GROUP NECG ANCOVA DESIGN

	Experimental (n = 23)	Control (n = 16)	Effect	P
General Knowledge	29.65	23.95	5.70	<.01
Numbers	34.84	26.72	8.12	<.01
Relations	6.21	4.19	2.02	<.01
Embedded Figures	21.47	19.64	1.83	<.05
Parts of the Whole	6.36	5.11	1.25	
Ability to Classify	8.69	5.57	3.12	<.01
Letters and Words	16.89	12.91	3.98	<.01
Classification Skills	14.09	11.44	2.65	<.01
Oral Comprehension	26.11	21.16	4.95	<.01

Estimators of the treatment effect on the midtest, adjusted for the pretest, were then calculated for the 'more able' experimental group NECG ANCOVA design, as shown in Table 8.

TABLE 8

MEANS FOR MIDTEST ADJUSTED BY COVARIANCE FOR PRETEST
IN THE 'MORE ABLE' EXPERIMENTAL GROUP NECG ANCOVA DESIGN

	Experimental (n = 23)	Control (n = 16)	Effect	P
General Knowledge	23.72	21.21	2.51	<.01
Numbers	27.72	24.16	3.56	<.01
Relations	5.58	4.53	1.05	<.01
Embedded Figures	18.78	17.19	1.59	<.05
Parts of the Whole	5.59	4.34	1.25	<.05
Ability to Classify	7.10	4.98	2.12	<.01
Letters and Words	14.59	9.91	4.68	<.01
Classification Skills	11.03	8.59	2.44	<.01
Oral Comprehension	21.31	17.05	4.26	<.01

To test whether an artifactual bias, overrating the experimental effect in 'more able' experimental group NECG ANCOVA design was present, a binomial test of null hypothesis #2 was performed (see Table 9).

TABLE 9

COMPARISON OF 'MORE ABLE' EXPERIMENTAL GROUP NECG ANCOVA DESIGN ESTIMATORS OF EFFECTS WITH UNBIASED EFFECTS ESTIMATORS

	Unbiased Estimator	NECG Design Estimator	Unbiased Estimator ≤ NECG Design Estimator
POSTTEST-PRETEST			
General Knowledge	5.13	5.70	yes
Numbers	7.01	8.12	yes
Relations	0.63	2.02	yes
Embedded Figures	1.76	1.83	yes
Parts of the Whole	0.80	1.25	yes
Ability to Classify	2.66	3.12	yes
Letters and Words	4.42	3.98	no
Classification Skills	2.33	2.65	yes
Oral Comprehension	3.08	4.95	yes
MIDTEST-PRETEST			
General Knowledge	0.43	2.51	yes
Numbers	2.70	3.56	yes
Relations	0.54	1.05	yes
Embedded Figures	1.11	1.59	yes
Parts of the Whole	0.79	1.25	yes
Ability to Classify	0.89	2.12	yes
Letters and Words	1.91	4.68	yes
Classification Skills	0.95	2.44	yes
Oral Comprehension	0.89	4.26	yes

Again the binomial test yeilded a $p < .001$ leading to a rejection of the null hypothesis and an acceptance of the alternate hypothesis that in 'more able' experimental group NECG ANCOVA designs a bias overrating the treatment effect appeared.

To answer the research question "Does the 'matching' NECG design produce an artifactual bias underrating whichever group is initially lower?", again the difference between the unbiased estimators of adjusted means from the ANCOVA design were employed as unbiased estimators of the experimental effect. Analyses of simulated NECG 'matching' designs were performed first with a 'less able' experimental group and then with a 'more able' experimental group, comparing in each analysis the experimental group with its matched control group to determine estimators of treatment effect in the NECG design.

For the 'less able' experimental group NECG design, posttest means for the experimental group and their matched 'more able' controls, the estimator of effect, and the significance of the difference between group posttest means (when $p < .05$) are given in Table 10.

TABLE 10

POSTTEST MEANS FOR THE 'LESS ABLE' EXPERIMENTAL GROUP AND FOR THE MATCHED CONTROL GROUP

	Experimental (n = 16)	Control (n = 16)	Effect	P
General Knowledge	28.25	25.94	2.31	
Numbers	32.75	25.94	6.81	<.05
Relations	5.06	5.62	-0.56	
Embedded Figures	21.06	19.50	1.56	
Parts of the Whole	6.00	5.56	0.44	
Ability to Classify	8.50	5.75	2.75	<.05
Letters and Words	17.06	11.31	5.75	<.01
Classification Skills	12.93	10.94	1.99	
Oral Comprehension	24.69	22.81	1.88	

Table 11 gives the midtest means for the 'less able' experimental group and their matched 'more able' controls. Also given in Table 11 are the estimators of treatment effect in this NECG design, and the significance of the difference of the midtest means between the two groups (when $p < .05$).

TABLE 11

MIDTEST MEANS FOR THE 'LESS ABLE' EXPERIMENTAL GROUP AND FOR THE MATCHED CONTROL GROUP

	Experimental (n = 16)	Control (n = 16)	Effect	P
General Knowledge	20.69	24.81	-4.12	<.05
Numbers	25.50	23.06	2.44	
Relations	4.38	4.38	0.00	
Embedded Figures	17.88	16.31	1.57	
Parts of the Whole	4.25	4.06	0.19	
Ability to Classify	5.94	5.69	0.25	
Letters and Words	13.56	12.75	0.81	
Classification Skills	8.31	9.50	-1.19	
Oral Comprehension	20.50	22.44	-1.94	

To test whether an artifactual bias, underrating the experimental effect in this 'less able' experimental group NECG 'matching' design was present, a binomial test of null hypothesis #3 was performed (see Table 12).

TABLE 12

COMPARISON OF 'LESS ABLE' EXPERIMENTAL GROUP NECG 'MATCHING' DESIGN ESTIMATORS OF EFFECTS WITH UNBIASED EFFECTS ESTIMATORS

	Unbiased Estimator	NECG Design Estimator	Unbiased Estimator > NECG Design Estimator
POSTTEST-PRETEST			
General Knowledge	5.13	2.31	yes
Numbers	7.01	6.81	yes
Relations	0.63	-0.56	yes
Embedded Figures	1.76	1.56	yes
Parts of the Whole	0.80	0.44	yes
Ability to Classify	2.66	2.75	no
Letters and Words	4.42	5.75	no
Classification Skills	2.33	1.09	yes
Oral Comprehension	3.08	1.88	yes
MIDTEST-PRETEST			
General Knowledge	0.43	-4.12	yes
Numbers	2.70	2.44	yes
Relations	0.54	0.00	yes
Embedded Figures	1.11	1.57	no
Parts of the Whole	0.79	0.19	yes
Ability to Classify	0.89	0.25	yes
Letters and Words	1.91	0.81	yes
Classification Skills	0.95	-1.19	yes
Oral Comprehension	0.89	-1.94	yes

The binomial test yielded a $p < .01$ leading to a rejection of the null hypothesis and an acceptance of the alternate hypothesis that in 'less able' experimental group NECG 'matching' designs, a bias underrating the treatment effect appeared.

To test whether artifactual bias was created in the 'more able' experimental group NECG 'matching' design, such a design was simulated by selecting the experimental group from the 'more able' center and their matches from the 'less able' center. Posttest means for this experimental group and for their 'less able' matches, as well as estimators of effect and the significance of the difference between group posttest means (when $p < .05$) are given in Table 13.

TABLE 13

POSTTEST MEANS FOR THE 'MORE ABLE' EXPERIMENTAL GROUP AND FOR THE MATCHED CONTROL GROUP

	Experimental (n = 23)	Control (n = 23)	Effect	P
General Knowledge	31.48	18.09	13.39	<.01
Numbers	35.83	21.48	14.35	<.01
Relations	6.22	4.09	2.13	<.01
Embedded Figures	21.59	18.23	3.36	<.05
Parts of the Whole	6.32	4.23	2.09	<.01
Ability to Classify	8.87	5.00	3.87	<.01
Letters and Words	16.87	11.70	5.17	<.01
Classification Skills	14.39	9.13	5.26	<.01
Oral Comprehension	26.44	17.17	9.27	<.01

Table 14 gives the midtest means for this 'more able' experimental group and for their matched 'less able' controls. Also given in Table 14 are the estimators of treatment effect in this NECG design and the significance of the difference between the midtest means of the two groups (when $p < .05$).

TABLE 14

MIDTEST MEANS FOR THE 'MORE ABLE' EXPERIMENTAL GROUP AND FOR THE MATCHED CONTROL GROUP

	Experimental (n = 23)	Control (n = 23)	Effect	P
General Knowledge	26.09	15.65	10.44	<.01
Numbers	28.87	19.39	9.48	<.01
Relations	5.78	4.09	1.69	<.01
Embedded Figures	19.00	15.90	3.10	
Parts of the Whole	5.56	3.74	1.82	<.01
Ability to Classify	7.27	4.77	2.50	<.01
Letters and Words	14.55	8.96	5.59	<.01
Classification Skills	11.44	7.52	3.92	<.01
Oral Comprehension	21.86	13.77	8.09	<.01

To test whether an artifactual bias, overrating the experimental effect in this 'more able' experimental group NECG 'matching' design was present, a binomial test of null hypothesis #4 was performed (see Table 15).

TABLE 15

COMPARISON OF 'MORE ABLE' EXPERIMENTAL GROUP NECG 'MATCHING' DESIGN ESTIMATORS OF EFFECTS WITH UNBIASED EFFECTS ESTIMATORS

	Unbiased Estimator	NECG Design Estimator	Unbiased Estimator < NECG Design Estimator
POSTTEST-PRETEST			
General Knowledge	5.13	13.39	yes
Numbers	7.01	14.35	yes
Relations	0.63	2.13	yes
Embedded Figures	1.76	3.36	yes
Parts of the Whole	0.80	2.09	yes
Ability to Classify	2.66	3.87	yes
Letters and Words	4.42	5.17	yes
Classification Skills	2.33	5.26	yes
Oral Comprehension	3.08	8.27	yes
MIDTEST-PRETEST			
General Knowledge	0.43	10.44	yes
Numbers	2.70	9.48	yes
Relations	0.54	1.69	yes
Embedded Figures	1.11	3.10	yes
Parts of the Whole	0.79	1.82	yes
Ability to Classify	0.89	2.50	yes
Letters and Words	1.91	5.59	yes
Classification Skills	0.95	3.92	yes
Oral Comprehension	0.89	8.09	yes

The binomial test yielded a $p < .001$ leading to a rejection of the null hypothesis and an acceptance of the alternate hypothesis that in 'more able' experimental group NECG 'matching' designs, a bias overrating the treatment effect appeared.

Statistical Significance of Treatment Effect Across Designs

In both the 'matching' and the covariance NECG designs with adjustment for pretest on either the posttest or the midtest, a bias in adjustment in the direction predicted by Campbell appeared in almost all instances. While significant biases appeared as predicted in each NECG design, perhaps the most interesting insight into the NECG quasiexperimental design bias was gained from an examination of the decisions regarding the statistical significance of the treatment effect across designs. Table 16 gives the levels of significance (when $p < .05$) obtained for the experimental effect in each design when the posttest was adjusted for the pretest.

TABLE 16

LEVELS OF SIGNIFICANCE OF THE TREATMENT EFFECT (<.05) ACROSS DESIGNS WHEN POSTTEST WAS ADJUSTED FOR PRETEST

		Covariance		Matching	
	Unbiased Estimator	'More Able' Exp. Group Estimator	'Less Able' Exp. Group Estimator	'More Able' Exp. Group Estimator	'Less Able' Exp. Group Estimator
Knowledge	<.01	<.01		<.01	
Numbers	<.01	<.01	<.05	<.01	<.05
Relations		<.01		<.01	
Figures	<.01	<.05		<.05	
Parts				<.01	
Ability	<.01	<.01		<.01	
Letters	<.01	<.01	<.01	<.01	<.01
Class. Sk.	<.01	<.01		<.01	
Oral Com.	<.01	<.01		<.01	

Both the 'matching' and covariance 'more able' experimental group designs yielded more instances of statistically significant treatment effects than did the true experimental design. This finding was consistent with Campbell's prediction. Also in support of Campbell was the fact that in both the 'matching' and covariance 'less able' experimental group designs, fewer treatment effects were detected than in the true experimental design. Had the original experiment employed any of these quasiexperimental designs, errors in the conclusions regarding the treatment effect, based purely on artifacts of the design employed, would have been made.

An examination of the derived significance of the treatment effect over the seven-week interval between the pretest and the midtest showed biases in designs other than the true experiment consistent with those found in the posttest-pretest examination. Table 17 presents levels of significance (when $p < .05$) obtained for the experimental effect in each design when the midtest was adjusted for the pretest.

TABLE 17

LEVELS OF SIGNIFICANCE OF THE TREATMENT EFFECT (<.05) ACROSS DESIGNS WHEN MIDTEST WAS ADJUSTED FOR PRETEST

		Covariance		Matching	
	Unbiased Estimator	'More Able' Exp. Group Estimator	'Less Able' Exp. Group Estimator	'More Able' Exp. Group Estimator	'Less Able' Exp. Group Estimator
Knowledge		<.01	<.01*	<.01	<.05*
Numbers	<.05	<.01		<.01	
Relations		<.01		<.01	
Figures		<.05			
Parts		<.05		<.01	
Ability		<.01		<.01	
Letters		<.01		<.01	
Class. Sk.		<.01		<.01	
Oral Com.		<.01		<.01	

* indicates a significant negative treatment effect

Again the 'matching' and covariance 'more able' experimental group designs both made the experimental group appear better than did the true experiment, as predicted by Campbell. Also in support of the prediction that regression artifacts would underrate the initially lower group was the fact that the 'less able' experimental group 'matching' and covariance designs made the experimental group appear worse than did the true experiment. Indeed, on the Knowledge test, the experimental group appeared to be significantly below the control group in both of these designs, while in the true experimental design there was a positive, but not significant, estimate of treatment effect.

CONCLUSIONS, RECOMMENDATIONS, AND IMPLICATIONS

The purpose of this study was to determine whether artifactual bias was introduced in NECG quasiexperimental ANCOVA and 'matching' designs. It also determined what differences in conclusions would have resulted had one of these latter designs been used in place of the true experimental design with this particular data set. Specifically it investigated whether differences in outcomes varied across designs so as to underrate whichever group was initially lower in both NECG designs. Data from a true, well-controlled experiment were used, and ANCOVA and 'matching' NECG designs were simulated using subsets of the data. Each design was tested in the case where the experimental group was 'more able' than the control group and in the case where the experimental group was 'less able' than the control group. The research hypothesis predicted that in each of the quasiexperimental designs, with a premeasure less than perfectly reliable, the quasiexperimental adjustment for initial intact group differences would underrate the lower group.

Firdings in this investigation indicated that in both quasiexperimental NECG designs a bias consistent with expectation was found. The findings from this investigation yield suggestions for future empirical investigations of regression artifacts in NECG designs. One might wish to investigate the effects of this bias with the use of traditional adjustment models both with and without regression homogeneity in groups. One might also wish to investigate more recent models, such as those proposed by Joreskog (1970, 1974), which allow for within-group correlated error.

Other possibilities for future study include the use of several true experimental data sets, each having strong treatment effects, to further substantiate whether the regression artifact does indeed appear across designs. Such a meta-analytic design would not be subject to the limitations imposed by the use of data from only one experiment. The use of data sets with larger group sizes would enhance the power of statistical tests employed and would lead to more conclusive findings than were obtainable in this study. Also, larger group sizes would allow for a systematic test of variations in power across designs. A final suggestion would be to operationalize the bias parameters as described by Cronbach et al. (1977) to verify empirically the conditions for overadjustment, underadjustment, and ideal adjustment when using covariance analysis.

The findings of this study lead to recommendations both for interpreting NECG design studies presently found in the literature and for the design of future studies. Interpretation of findings from studies employing NECG covariance or 'matching' designs must be with due caution. Statements supporting treatments when the experimental group was initially higher than the control group or those rejecting treatments when the experimental group was initially lower than the control group are particularly problematic.

With respect to future studies, the findings of this investigation add strength to conclusions given by Campbell and Erlebacher (1970) and Campbell and Boruch (1971), which stress the importance of random assignment of subjects to groups and of groups to treatments when using traditional methodologies to "control" for premeasure differences. Cronbach et al. (1977) recommend that "the investigator who wishes to make adjustments when assignment is not controlled will usually be wise to take many covariates into his analysis" (Cronbach et al., 1977, p.21). This author concurs with these suggestions and also suggests the possible use of nontraditional regression models which avoid this regression artifact when NECG designs are necessary. Among these alternative models are latent structure models, as proposed by Joreskog (1970, 1974), which can be specified to allow for correlated error among subjects within groups. Such models are being used more frequently in recent studies (see, for example, Magidson, 1977). Other alternative models are those which use no control group, but rather predict outcome without a treatment and then examine outcomes, given the treatment, to see if effects are present (e.g., Kenny, 1975, Bryk, 1977, Bryk & Weisberg, 1977, Kenny & Cohen, 1979). Care must be taken, however, that limitations and biases implicit in these alternative models be fully investigated before their acceptance as the solution to problems of traditional models. It is recommended, based on both the literature and on the findings in this study, that an effort be made to randomize in any study, regardless of the design and models employed, to ensure maximum

generalizability of findings.

REFERENCES

1. Ball, S., and Bogatz, G.A., "Research on Sesame Street: Some Implications for Compensatory Education", Proceedings of the Second Annual Hyman Blumberg Symposium on Research in Early Childhood Education, 1972. Princeton: Johns Hopkins Press, Educational Testing Service.

2. Barnow, S., Bias in the Estimates of Treatment Effects in Quasi-Experimental Evaluations, May, 1976. Washington: ASPER-Office of Research and Evaluation, U.S. Department of Labor.

3. Bryk, A.S., An Investigation of the Effectiveness of Alternative Statistical Adjustment Strategies in the Analysis of Quasi-Experimental Growth Data, 1977. Thesis for the Ed.D.: Harvard University.

4. ____, and Weisberg, H.I., "Use of the Non-Equivalent Control Group Design When Subjects are Growing", 1977, Psychological Bulletin, Vol. 85, No.4.

5. Campbell, D.T., "Temporal Changes in Treatment-Effect Correlations: A Quasi-Experimental Model for Institutional Records and Longitudinal Studies", in G.V. Glass (ed.), Proceedings of the 1970 Invitational Conference on Testing Problems, 1971. Princeton: Educational Testing Service.

6. _____, and Boruch, R.F., "Making the Case for Randomized Assignment to Treatments by Considering the Alternatives: Six Ways in Which Quasi-Experimental Evaluations in Compensatory Education Tend to Underestimate Effects", in C.A. Bennett and A.A. Lumsdaine (eds.), Evaluation and Experiment: Some Critical Issues in Assessing Social Programs, 1975. Seattle, Washington: Academic Press, Inc.

7. _____, and Erlebacher, A., "How Regression Artifacts in Quasi-Experimental Evaluations Can Mistakenly Make Compensatory Education Look Harmful", in J. Hellmuth (ed.), Compensatory Education: A National Debate, Vol. III of The Disadvantaged Child, 1970. New York: Brunner/Mazel (pp. 185-210, Reply to the Replies, pp. 221-5).

8. _____, and Stanley, J.C., "Experimental and Quasi-Experimental Designs for Research on Teaching", in N.L. Gage (ed.), Handbook of Research on Teaching, 1963. Chicago: Rand McNally & Co. (pp. 171-246).

9. Cronbach, J., and Furby, L., "How We Should Measure Change--Or Should We?", 1970, Psychological Bulletin, Vol. 74, No. 1 (pp. 68-80).

10. _____, Rogosa, D.R., Floden, R.E., and Price, G.G., Analysis of Covariance: Angel of Salvation, or Temptress and Deluder?, 1976. Stanford, Ca.: Stanford University (Occasional papers of the Stanford Evaluation Consortium).

11. _____, Analysis of Covariance in Nonrandomized Experiments: Parameters Affecting Bias, 1977. Stanford, Ca.: Stanford University (Occasional papers of the Stanford Evaluation Consortium).

12. Cochran, W.G., "Matching in Analytical Studies", 1953, American Journal of Public Health, Part I (pp. 684-91).

13. _____, and Rubin, D.B., "Controlling Bias in Observational Studies: A Review", 1973, Sankhya, Series A., Vol. 35, Part 4 (pp. 417-46).

14. DeGracie, J.S., and Fuller, W.A., "Estimation of the Slope and Analysis of Covariance When the Concomitant Variable is Measured With Error", 1972, Journal of the American Statistical Association, Vol. 67 (pp. 930-7).

15. Diaz-Guerrero, R., Reyes-Lagunes, I., Witzke, D.B., and Holtzman, W., "Plaza Sesamo in Mexico: An Evaluation", 1976, Spring, Journal of Communication, Vol. 26, No. 2 (pp. 145-54).

16. Diaz-Guerrero, R., and Holtzman, W.H., "Learning by Televised 'Plaza Sesamo' in Mexico", 1974, Journal of Educational Psychology, Vol. 66, No. 5 (pp. 632-43).

17. _____, Learning by Televised Plaza Sesamo in Mexico, 1973. Mexico City: Centro de Investigaciones Psicopedagogicas, A.C. (Paper reporting results of study supported in part by Ford Foundation Grant No. 730-0206).

18. Director, S.M., Underadjustment Bias in the Quasi-Experimental Evaluation of Manpower Training, 1974, unpublished doctoral dissertation: Northwestern University.

19. Furby, L., "Interpreting Regression toward the Mean in Developmental Research", 1973, Developmental Psychology, Vol. 8, No. 2 (pp. 172-9).

20. Goldberger, A.S., "Selection Bias in Evaluating Treatment Effects: Some Formal Illustrations", Institute for Research on Poverty Discussion Papers, 1972. Madison, Wisconsin: University of Wisconsin.

21. Goodwin, W.L., and Sanders, J.R., The Use of Experimental and Quasi-Experimental Designs in Educational Evaluation (Research Report), 1972. Boulder, Colorado: University of Colorado, Laboratory of Educational Research.

22. Gordon, R.A., "Issues in Multiple Regression", 1973, American Journal of Sociology, February (pp. 592-616).

23. Joreskog, K.G., "Analyzing Psychological Data by Structural Analysis of Covariance Matrices", in D.H. Drantz, R.C. Atkinson, R.D. Luce, and P. Suppes (eds.), Contemporary Development in Mathematical Psychology, 1974. San Francisco, Ca.: W.H. Freeman, Vol. II.

24. _____, "A General Method for the Analysis of Covariance Structures", 1970, Biometrika, Vol. 57 (pp. 239-51).

25. Kenny, D.A., "A Quasi-Experimental Approach to Assessing Treatment Effects in the Non-Equivalent Control Group Design", 1975, Psycological Bulletin, Vol. 82, No. 3 (pp. 345-62).

26. _____, and Cohen, S.H., "Analysis of the Non-Equivalent Control Group Design under Sociological Selection", 1979, unpublished manuscript.

27. Lord, F.M., "Large-Scale Covariance Analysis when the Control Variable is Fallible", 1960, Journal of the American Statistical Association, Vol. 55 (pp. 307-21).

28. Magidson, J., "Toward a Causal Model Approach for Adjusting for Preexisting Differences in the Nonequivalent Control Group Situation: A General Alternative to ANCOVA", 1977, Evaluation Quarterly, Vol. 1, No. 3, August (pp. 399-420).

29. _____, Barnow, B., and Campbell, D., Correcting the Underadjustment Bias in the Original Head Start Evaluation, 1976, Evaluation Research Report No. 2JM, Evanston, Il.: Northwestern University Psychology Department.

30. McDill, E.L., McDill, M.S., and Sprehe, J., Strategies for Success in Compensatory Education: An Appraisal of Evaluation Research, 1969. Baltimore, Md.: Johns Hopkins Press.

31. McNemar, Q., "A Critical Examination of the University of Iowa Studies of Environmental Influences Upon the I.Q.", 1940, Psychological Bulletin, Vol. 37 (pp. 63-92).

32. O'Connor, E.F., Jr., Unraveling Lord's Paradox: The Appropriate Use of Multiple Regression Analysis in Quasi-Experimental Research, 1973. Princeton, New Jersey: Educational Testing Service (ETS RB 73-53).

33. Porter, A.C., The Effects of Using Fallible Variables in the Analysis of Covariance, June, 1967, Ph.D. dissertation: University of Wisconsin.

34. _____, Comments on Some Current Strategies to Evaluate the Effectiveness of Compensatory Education Programs and Comments on the Westinghouse-Ohio University Study, August, 1969, two memoranda prepared for Robert D. Hess for use at the symposium on "The Effectiveness of Contemporary Education Programs in the Early Years: Reports from Three National Evaluations and Longitudinal Studies", Washington, D.C., Annual Meeting of the American Psychological Association, August 31, 1969.

35. _____, and Chibucos, T.R., "Selecting Analysis Strategies", in G.D. Borich (ed.), Evaluating Educational Programs and Products, 1974. Englewood Cliffs, New Jersey: Educational Technology Publications.

36. Rulon, P.J., "Problems of Regression", 1941, Harvard Educational Review, Vol. 11 (pp. 213-23).

37. Reichardt, C.S., "The Statistical Analysis of Data from the Nonequivalent Control Group Design", in T.D. Cook and D.T. Campbell (eds.), The Design and Conduct of Quasi-Experiments and True Experiments in Field Settings, 1976. Chicago: Rand-McNally.

38. Rindskopf, D.M., A Comparison of Various Regression-Correlation Methods for Evaluating Nonexperimental Research, 1976. Ames, Iowa: Ph.D. dissertation, Iowa State University.

39. Sechrest, L., "Incremental Validity: A Recommendation", Educational and Psychological Measurement, 1963, Vol. 23 (pp. 153-8).

40. Thorndike, R.L., "Regression Fallacies in the Matched Groups Experiment", 1942, Psychometrika, Vol. 7 (pp. 85-102).

41. Vermillion, J.E., An Empirical Test of the Theory of Artifacts of Regression using Matching, Analysis of Covariance, and Raw Gain Scores with Non-Equivalent Control Groups in Quasi-Experimental Designs, 1979, Ph.D. dissertation: University of Virginia. University Microfilms, Ann Arbor, Michigan.

42. Werts, C.E., and Linn, R.L., "Lord's Paradox: A Generic Problem", 1969, Psychological Bulletin, Vol. 72 (pp. 423-5).

43. _____, Analysing School Effects: ANCOVA With a Fallible Covariate, July, 1969. Princeton, N.J.: Educational Testing Service Research Bulletin No. 69-59.

44. _____, Rock, D.A., and Joreskog, K.G., "Comparison of Correlations, Variances, Covariances, and Regression Weights with or without Measurement Error", 1976, Psychological Bulletin, Vol 83, No. 6 (pp. 1307-13).

45. Wolins, L., "The Use of Multiple Regression Procedure When the Predictor Variables are Psychological Tests", 1967, Educational and Psychological Measurement, Vol. 27 (pp. 821-7).